Published online 22 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D1235–D1240***
doi:10.1093/nar/gkr962

DNA*traffic*—a new database for systems biology of DNA dynamics during the cell life

Krzysztof Kuchta[1], Daniela Barszcz[1], Elzbieta Grzesiuk[1], Pawel Pomorski[2] and Joanna Krwawicz[1,*]

[1]Department of Molecular Biology, Institute of Biochemistry and Biophysics, Polish Academy of Sciences and [2]Department of Biochemistry, Nencki Institute of Experimental Biology, Warsaw, Poland

Received August 16, 2011; Revised October 13, 2011; Accepted October 14, 2011

ABSTRACT

DNA*traffic* (http://dnatraffic.ibb.waw.pl/) is dedicated to be a unique comprehensive and richly annotated database of genome dynamics during the cell life. It contains extensive data on the nomenclature, ontology, structure and function of proteins related to the DNA integrity mechanisms such as chromatin remodeling, histone modifications, DNA repair and damage response from eight organisms: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Escherichia coli* and *Arabidopsis thaliana*. DNA*traffic* contains comprehensive information on the diseases related to the assembled human proteins. DNA*traffic* is richly annotated in the systemic information on the nomenclature, chemistry and structure of DNA damage and their sources, including environmental agents or commonly used drugs targeting nucleic acids and/or proteins involved in the maintenance of genome stability. One of the DNA*traffic* database aim is to create the first platform of the combinatorial complexity of DNA network analysis. Database includes illustrations of pathways, damage, proteins and drugs. Since DNA*traffic* is designed to cover a broad spectrum of scientific disciplines, it has to be extensively linked to numerous external data sources. Our database represents the result of the manual annotation work aimed at making the DNA*traffic* much more useful for a wide range of systems biology applications.

INTRODUCTION

A comprehensive understanding of the maintenance of DNA integrity during the cell life requires the thorough characterization of many simple data concerning all nuclear processes involving DNA, and including replication, repair, recombination (3R) and transcription. The major processes that regulate chromatin structure and counterbalance its repressive effects are: (i) chromatin remodeling, (ii) post-translational histone modifications and (iii) histone replacement. Chromatin is a dynamic structure that modulates the access of regulatory factors to the genetic material. The main role of DNA molecules is the long-term storage of information, genetic instruction used in the development and functioning of all known living organisms (with the exception of RNA viruses). Cells are continuously exposed to damaging agents whose action results in modification of nucleic acids. DNA damage from endogenous sources gives rise to 20 000 lesions/mammalian cell/day (1). Lesions are also caused by errors in DNA metabolic processes, including the formation of single and double-strand breaks from the collapse of replication forks and the introduction of modified nucleic acid bases during DNA replication. Counting all together, daily the 10^{16}–10^{18} repair events occur in a healthy adult man (10^{12} cells) (2). On the other hand, DNA damage is also caused by the environmental factors such as chemicals, UV light and ionizing radiation. Also, DNA structure and some proteins involved in DNA replication and repair are targets for the drugs used during chemotherapy (3). The available anticancer drugs have distinct mechanisms of action, which may vary in their effects on different types of normal and cancer cells. Their role is to slow and hopefully halt the growth and spread of a cancer.

Across the evolutionary spectrum, living organisms depend on high-fidelity DNA replication and recombination mechanisms have to response to DNA damage and balance between the harmful and beneficial effects of manipulation into the genetic code. The knowledge of the processes in charge of DNA metabolism is critical to our understanding of how and why the genome is affected during the lifespan of the organism, and how the DNA repair systems efficiently work via several different pathways to protect the genome from potential mutagenic

*To whom correspondence should be addressed. Tel/Fax: +482 2592 3337; Email: joanna.krwawicz@gmail.com

modification and allow accurate transmission of genetic information (4). Unrepaired lesions or strand brakes left in DNA might be the result of dysfunction in DNA repair, and lead to aging, carcinogenesis or neurodegeneration (5,6). Some pathological disorders are directly related to defects in DNA repair, telomere maintenance or DNA damage response machinery (7–9). At the same time, random changes in DNA are viewed as a main source of genetic variability (e.g. the antibody production), and thus a driving force for evolution. A precise coordination of the genome networks is crucial to ensure that the correct genetic code is maintained within the genome.

Traditionally, most of the known information on the DNA study, DNA damage, diseases and drug targets has been resisted in books, journals and databases. Moreover, research in molecular biology has been focused on single problems, simplified to the maximal extend. Recently, the holistic approach to research, referred to as the systems biology, has gained importance and interest in the scientific community. Arising databases mainly contain information on the sequenced genomes, genes, proteins, RNAs, etc. Also, the depositaries of information on drugs, small molecules and chemicals are in common use. The topic of the DNA metabolism is covered by many computational resources. Metabolic pathway databases contain metabolic pathways from a wide variety of organisms (10–17). Those databases queried about 'DNA metabolism', 'DNA replication', 'DNA repair', 'nucleic acid' show several dozen answers. However, the chromatin maintenance network contains about 20 subpathways, depending on the organism [e.g. *Escherichia coli* cells lack of non-homologous end joining (NHEJ) repair or Fanconi anemia (FA) pathway].

In contrast to others, the open access DNA*traffic* database is a richly annotated resource for systems biology of DNA research containing information on: (i) DNA metabolism (replication, transcription, DNA repair pathways, chromatin organization, histone modifications and the DNA damage response network in eukaryotic and prokaryotic organisms); (ii) proteins enrolled in widely understanding the DNA metabolism; (iii) DNA damage (damage type, damage source and damage effect); (iv) diseases related to the assembled human proteins and (v) drugs targeted on nucleic acids metabolism and proteins involved in the maintenance of genome stability.

DNA*traffic* database for systems biology of genome integrity is addressed to scientists, pharmacologists and students.

DETAILS RELATING TO DNA*traffic*'s OVERALL DESIGN AND DATA STRUCTURE DEPICTION CONVENTIONS

The aspects of the biochemistry and molecular biology of the genome dynamics during the cell life are the key for learning genome stability networks. During DNA replication, transcription and DNA repair, the cellular machineries performing these tasks need to gain access to the DNA that is packaged into chromatin or nucleoid. The main aim of the DNA*traffic* database is to cover and elucidate the interdisciplinary knowledge linking all aspects of the DNA integrity processes (e.g. chromatin dynamics, DNA replication, damage signaling and DNA repair), DNA damage and drugs interacting with DNA or proteins directly enrolled in DNA metabolism and connect all pieces together for the coordination of steps within a pathway or for crosstalk between different pathways. As transcription, recombination and DNA integrity are central components in the evolution of recent genome structures, and because replication, recombination and repair (3R) were fundamental prerequisites for the origin of life, all these topics are taken under analysis and serve as the cohesive force underlying this comprehensive DNA topic-focused database (18).

PathCARD

We used KEGG (13) and Reactome (12) databases for data implementation about pathways and networks concerning DNA metabolism. Some data like prokaryotic SOS response and translesion synthesis (TLS) were directly added by our DNA*traffic* team. All proteins are classified according to the orthology class, and next to the DNA integrity networks: chromatin organization and histone modifications, replication, damage checkpoint, DNA repair, modulation of nucleotide pools and so on (Table 1). It must be emphasized that all described processes are tightly connected to each other and they act in concert sharing some steps and/or proteins. Known functions of proteins are indicated in the curator comments section of each entry. A special emphasis is devoted to the function of that protein within DNA metabolism pathways but we also refer to alternative roles in other pathways. Additionally, all Gene Ontology terms associated to that protein are listed. The pathway in which a given protein is playing a role is also explored by linking from DNA*traffic* to the pathways included in the KEGG and Reactome databases.

ProteinCARD

According to the DNA metabolism network we used the UniProt (19), KEGG (13) and National Center for Biotechnology Information (NCBI) databases for protein data implementation into DNA*traffic* database for eight model organisms commonly used for DNA study. We collected 2921 proteins, for example—582 for *Homo sapiens*, 277 for S*accharomyces cerevisiae* and 91 for *E. coli* (as of 13 October 2011). Using direct access from DNA*traffic* to protein all users can obtain unusual view of well-known proteins from model organisms but classified into the orthology classes. This innovation may be useful for the systems biology research and proper selection of the model organism for further study (Figure 1) of selected pathway. Amino acids and DNA sequences were downloaded from Ensemble. When available, links to the protein 3D structure in Protein Data Bank (PDB) were provided and 2D picture is visible in the single ProteinCARD entry. If annotated, possible physical interactions with other proteins were obtained through IntAct, STRING and other databases providing interacting

Table 1. Distribution of the orthology classes into DNA maintenance network in DNA*traffic* database

General name of super-pathway	Name of process	Number of orthology class
Genome dynamics	Chromatin remodeling	72
	Histones	5
	Histone modifications	75
	Transcription	34
	Transcription factors	35
	Heterochromatin formation	27
	Telomere maintenance	23
	Prokaryotic nucleoid remodeling	12
DNA synthesis	DNA replication	67
	TLS	13
	Rad6 pathway	5
DNA repair	Direct repair	
	Direct reversal	8
	Single-strand breaks repair	
	Base excision repair	39
	Nucleotide excision repair	47
	Mismatch repair	42
	Very short patch repair	1
	Double-strand breaks repair	
	Homologous recombination	56
	FA pathway	58
	NHEJ	17
Damage checkpoint	DDR	23
	Prokaryotic SOS response	18
Nucleotide level	Modulation of nucleotide pools	6

protein pairs from small and large-scale experiments. Also, manual annotation work was needed to match the DNA damage or drug to appropriate protein and DNA structure. Also, the DNA metabolism-related proteins from the DNA*traffic* database were classified by orthology into the functional (or predicted by orthology) activities such as DNA polymerase, DNA ligase, DNA glycosylase, DNA helicase, nuclease, etc. This action also needed manual annotation work. Using our knowledge and bioinformatics tools, in near future, protein will be classified into the structural families attending to the presence of characteristic domains, e.g. BRCA1, BARD1, BRCT and RING, etc (20,21). Each protein possesses its own ProteinCARD entry with a succinct description, reciprocal links to pathway(s), and if existing—to disease and DNA damage, and additional external links to NCBI, KEGG and UniProt databases. Moreover, from single ProteinCARD user can overview and access to the other proteins from the same orthology class. This information can be useful for the systemic study of DNA integrity.

DiseaseCARD

Disease is a condition, which arises in a living organism, animal or plant, when something malfunctions and impairs the normal operation of the organism. Developing an understanding of the factors that cause disease motivates most of biological research. Till now, DNA*traffic* collects 121 diseases related to dysfunction in 77 proteins enrolled to the DNA networks. Data were implemented from OMIM (22) and KEGG databases (13) as well as directly from PubMed. This action needed manual annotation work. Each disease possesses its own DiseaseCARD entry with a succinct description, link to protein(s) and sometimes the picture of the symptoms. Reciprocal links to diseases are also available in each protein and pathway field (Figure 1).

DamageCARD

As of 13 October 2011, we collected information about 146 different types of damage in the DNA. Many of them describe general classes of damage events such as methylation or oxidative damage, or single-strand breaks or base loss, which are independent of the local sequence. About 50 chemical compounds that cause DNA damage were connected to the appropriate types of damage. Each type of damage is described on its own DamageCARD entry that includes information about the potential source (e.g. spontaneous formation, intermediate in some DNA repair process, methylating agents, etc.), proteins that may recognize its presence in the DNA, keywords that facilitate analyzing its context and external links (if available) to: PubChem Compound (CID), PubChem Substance (SID), ChemSpider, KEGG Compound, ChEBI and ChEMBL. DNA*traffic* database also displays the unique chemical structures of DNA lesions in 2D and provides atomic coordinates for download in the smiles, InChi and InChiKey format.

DrugCARD

Till now, we collected information about over 181 different types of drugs interacting with DNA or proteins involved in nucleic acids metabolism. Data were implemented from DrugBank, T3DB, Therapeutic Target Database (TTD), KEGG Compounds databases (13,23–25). Each type of drug is described on its own DrugCARD entry that includes information about the potential application (e.g. anticancer treatment, DNA topoisomerase inhibitor and other), drug–protein or drug–DNA interaction and external links to DrugBank, KEGG Compound, PubChemCompound, PubChem Substance, ChemSpider, ChEBI, ChEMBL and TTD databases. DNA*traffic* database also displays the unique chemical structures of drugs in 2D and provides atomic coordinates for download in the smiles, InChi and InChiKey format.

SCHEME OF THE DNA*traffic* DATABASE ARCHITECTURE

The unordered data are difficult to interpret and many of the connections are lost. The OWT ontology provides the clear view and discovers the new connections. DNA*traffic* database has been implemented using the Django web framework (http://www.djangoproject.com/). It uses a PostgreSQL relational database to store data (http://www.postgresql.org/). Scripts are written in Phyton language. DNA*traffic* database is freely available and can be accessed at http://dnatraffic.ibb.waw.pl/dnatraffic/.

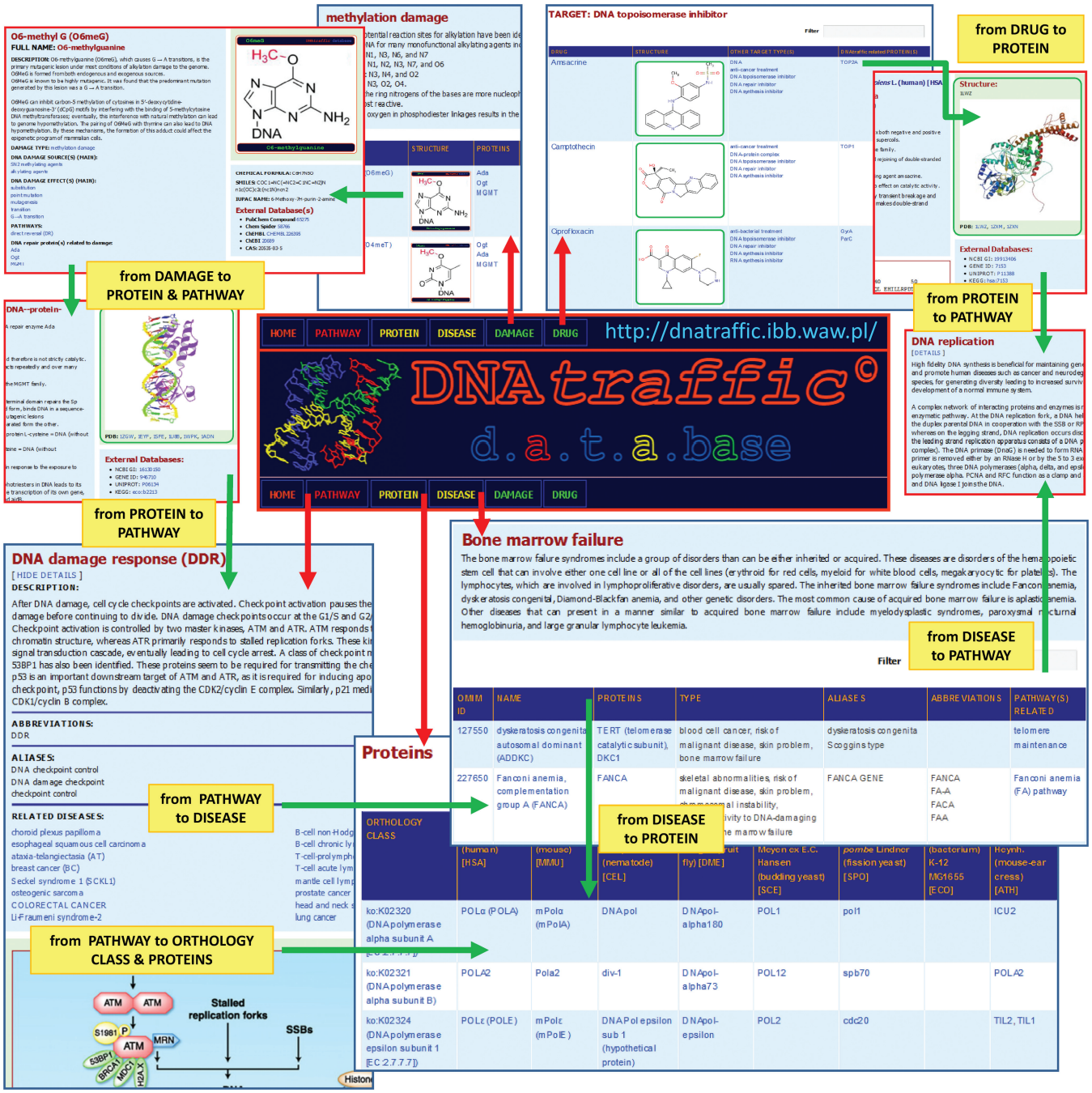

Figure 1. Scheme of DNA*traffic* database.

EXPANDED DATABASE LINKAGES

Because DNA*traffic* was designed to cover a broad spectrum of scientific disciplines, it must be extensively linked to many external databases. Until now, DNA*traffic* contains up to 15 database hyperlinks including links to KEGG (13), UniProt (19), OMIM (22), PDB (26), PubChem (27), ChEBI (28), ChEMBL, GenBank (29), Pfam (30), GeneCards (31), GenAtlas (32), HGNC, PubMed, ChemSpider (33) and TTD (25).

CONCLUSION

Researchers of the various chromatin structure and DNA repair processes have recently embraced approaches in which global measurements of gene expression and the proteome can be combined with genome-wide screening of sensitivity mutants to develop an integrated view of how cells respond to and protect themselves against DNA damaging agents. The emerging picture from these global genomic studies is quite different from the previous concept of DNA repair, cell cycle control and induction of

apoptosis as being independent processes. In fact these processes appear to form a fully integrated network. Integration of these genome-wide measurements allows the development of specific models of response networks that could not have been detected or discerned previously.

DNA*traffic* database is the first platform for systems biology of DNA integrity during the cell life, and can be also integrally involved in translational research (18). This includes the identification of small molecule inhibitors of novel DNA damage response (DDR) pathways that put new light on the causes of cancer or have potential uses in treatment.

DNA*traffic* contains a significant number of data. As highlighted throughout this article, numerous improvements have been made in the quantity, quality, depth and organization of the information provided. DNA*traffic* contains illustrated DNA networks in the cell, protein, damage and drug structures data and pictures. DNA*traffic* also offers expanded database links. It is hoped that DNA*traffic* will continue to develop to fulfil the needs of its users and provide an increasingly useful, information-rich DNA metabolism resource.

ACKNOWLEDGEMENTS

The authors would like to thank Professor Jaroslaw Kusmierak (Institute of Biochemistry and Biophysics, Polish Academy of Sciences) for his help with the DNA damage nomenclature and chemical structure instructions drawing.

FUNDING

A grant from the Norwegian Financial Mechanism (PNRF-143-AI-1/07, in part); Polish Ministry of Science and Higher Education (N N301 165835, in part). Funding for open access charge: partially waived by Oxford University Press (55%), and Institute of Biochemistry and Biophysics PAS, Warsaw, Poland (45%).

Conflict of interest statement. None declared.

REFERENCES

1. Lindahl,T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
2. Friedberg,E.C., Aguilera,A., Gellert,M., Hanawalt,P.C., Hays,J.B., Lehmann,A.R., Lindahl,T., Lowndes,N., Sarasin,A. and Wood,R.D. (2006) DNA repair: from molecular mechanism to human disease. *DNA Repair*, **5**, 986–996.
3. Krwawicz,J. (2011) DNA damage and repair in radio- and chemotherapy. *Acta Biochim. Pol.*, **58**, 433–442.
4. Kunkel,T.A. (2009) Evolving views of DNA replication (in)fidelity. *Cold Spring Harb. Symp. Quant. Biol.*, **74**, 91–101.
5. Reynolds,J.J., El-Khamisy,S.F. and Caldecott,K.W. (2009) Short-patch single-strand break repair in ataxia oculomotor apraxia-1. *Biochem. Soc. Trans.*, **37**, 577–581.
6. Jeggo,P. The role of the DNA damage response mechanisms after low-dose radiation exposure and a consideration of potentially sensitive individuals. *Radiat. Res.*, **174**, 825–832.
7. Arczewska,K.D. and Kusmierek,J.T. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 2. role of bacterial mutator gene homologues in human disease. Overview of nucleotide pool sanitization and mismatch repair systems. *Acta Biochim. Pol.*, **54**, 435–457.
8. Krwawicz,J., Arczewska,K.D., Speina,E., Maciejewska,A. and Grzesiuk,E. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 1. mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease. *Acta Biochim. Pol.*, **54**, 413–434.
9. Maddukuri,L., Dudzinska,D. and Tudek,B. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 4. The role of nucleotide excision DNA repair (NER) system in mammalian cells. *Acta Biochim. Pol.*, **54**, 469–482.
10. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
11. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
12. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
13. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
14. Milanowska,K., Krwawicz,J., Papaj,G., Kosinski,J., Poleszak,K., Lesiak,J., Osinska,E., Rother,K. and Bujnicki,J.M. (2011) REPAIRtoire - a database of DNA repair pathways. *Nucleic Acids Res.*, **39**, D788–D792.
15. Podlevsky,J.D., Bley,C.J., Omana,R.V., Qi,X. and Chen,J.J. (2008) The telomerase database. *Nucleic Acids Res.*, **36**, D339–D343.
16. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE - a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
17. Romero,P., Wagg,J., Green,M.L., Kaiser,D., Krummenacker,M. and Karp,P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.
18. Krwawicz,J. (2011) Integration of knowledge from open-access databases and usage bioinformatics tools in DNA network database creation. *Acta Biochim. Pol.*, **58**, 455–462.
19. UniProt. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
20. Kuchta,K., Knizewski,L., Wyrwicz,L.S., Rychlewski,L. and Ginalski,K. (2009) Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. *Nucleic Acids Res.*, **37**, 7701–7714.
21. Aravind,L., Walker,D.R. and Koonin,E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.*, **27**, 1223–1242.
22. McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
23. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2010) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
24. Lim,E., Pon,A., Djoumbou,Y., Knox,C., Shrivastava,S., Guo,A.C., Neveu,V. and Wishart,D.S. (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.*, **38**, D781–D786.
25. Zhu,F., Han,B., Kumar,P., Liu,X., Ma,X., Wei,X., Huang,L., Guo,Y., Han,L., Zheng,C. *et al.* (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **38**, D787–D791.
26. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.

27. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
28. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
29. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
30. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
31. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
32. Roux-Rouquie,M., Chauvet,M.L., Munnich,A. and Frezal,J. (1999) Human genes involved in chromatin remodeling in transcription initiation, and associated diseases: An overview using the GENATLAS database. *Mol. Genet. Metab.*, **67**, 261–277.
33. Williams,A.J. (2008) Public chemical compound databases. *Curr. Opin. Drug Discov Dev.*, **11**, 393–404.